(19)日本国特許庁.(JP)　(12)公開特許公報(A)　(11)特許出願公開番号

特開平11－184890

(43)公開日　平成11年(1999)7月9日

(51)Int.Cl.[6]	識別記号	FI		
G06F 17/30		G06F 15/401	320A	
17/27		15/38	D	

審査請求　未請求　請求項の数3　FD　(全 9 頁)

(21)出願番号　特願平9－364535

(22)出願日　平成9年(1997)12月18日

(71)出願人　000005496
富士ゼロックス株式会社
東京都港区赤坂二丁目17番22号

(72)発明者　岡　満美子
神奈川県足柄上郡中井町境430　グリーンテクなかい　富士ゼロックス株式会社内

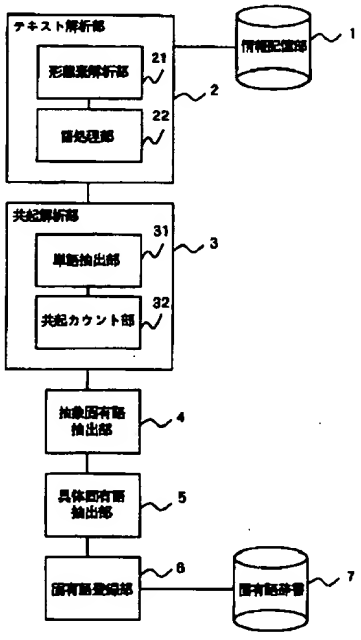
(74)代理人　弁理士　守山　辰雄

(54)【発明の名称】　個人関心事辞書作成装置

(57)【要約】

【課題】　個人が関心をもつ知識領域を表す語句（固有語）を当該個人が書いたテキストデータから抽出して、当該個人の関心事を把握するための辞書を作成する。

【解決手段】　情報記憶手段1に記憶されている特定個人によって作成されたテキストデータを含む文書を、テキスト解析手段2で形態素解析等して語句の単位に切り分け、共起解析手段3によって、テキストデータ中の予め定めたテキスト単位中で固有名詞と共起する語を当該テキストデータを作成した特定個人の知識領域を表す抽象固有語候補として抽出し、更に、当該固有名詞を抽出して各抽象固有語との共起頻度をカウントする。そして、この固有名詞の共起頻度を基準として、抽象固有語抽出手段4で抽象固有語候補から抽象固有語を抽出し、更に、具体固有語抽出手段5により、抽出された抽象固有語と共起する固有名詞を、当該抽象固有語が表す領域での前記特定個人の具体的な関心を表す具体固有語として抽出し、これら抽出された抽象固有語と具体固有語とを対応付けて固有語辞書手段7に保持する。



1

【特許請求の範囲】

【請求項１】　文書中から当該文書を作成した特定個人の具体的な関心を表す語を抽出して、辞書にまとめる個人関心事辞書作成装置であって、

特定個人によって作成されたテキストデータを含む文書を記憶する情報記憶手段と、

前記文書中のテキストデータを解析するテキスト解析手段と、

前記テキストデータ中の予め定めたテキスト単位中で固有名詞と共起する語を当該テキストデータを作成した特定個人の知識領域を表す抽象固有語候補として抽出し、更に、当該固有名詞を抽出して各抽象固有語との共起頻度をカウントする共起解析手段と、

前記共起する固有名詞の共起頻度を基準として前記抽象固有語候補から抽象固有語を抽出する抽象固有語抽出手段と、

前記抽出された抽象固有語と共起する固有名詞を、当該抽象固有語が表す領域での前記特定個人の具体的な関心を表す具体固有語として抽出する具体固有語抽出手段と、

前記抽出された抽象固有語と具体固有語とを対応付けて保持する固有語辞書手段と、

を備えたことを特徴とする個人関心事辞書作成装置。

【請求項２】　文書中から当該文書を作成した特定個人の具体的な関心を表す語を抽出して、辞書にまとめる個人関心事辞書作成装置であって、

特定個人によって作成されたテキストデータを含む文書を記憶する情報記憶手段と、

前記文書中のテキストデータを解析するテキスト解析手段と、

前記テキストデータ中の予め定めたテキスト単位中で固有名詞と共起する語を当該テキストデータを作成した特定個人の知識領域を表す抽象固有語候補として抽出し、更に、当該固有名詞を抽出して個数をカウントする共起解析手段と、

前記抽出された固有名詞数を基準として前記抽象固有語候補から抽象固有語を抽出する抽象固有語抽出手段と、

前記抽出された抽象固有語と共起する固有名詞を、当該抽象固有語が表す領域での前記特定個人の具体的な関心を表す具体固有語として抽出する具体固有語抽出手段と、

前記抽出された抽象固有語と具体固有語とを対応付けて保持する固有語辞書手段と、

を備えたことを特徴とする個人関心事辞書作成装置。

【請求項３】　請求項１又は請求項２に記載の個人関心事辞書作成装置において、

前記情報記憶手段は同一の個人によって作成された複数の文書を記憶しており、

前記テキスト解析手段、前記共起解析手段、前記抽象固有語抽出手段、及び、前記具体固有語抽出手段は、同一の個人に係る複数の文書に対して前記処理を施し、

前記固有語辞書手段には前記複数の文書中から抽出した同一の個人に係る抽象固有語と具体固有語とが対応付けて保持されることを特徴とする個人関心事辞書作成装置。

2

【発明の詳細な説明】

【０００１】

【発明の属する技術の分野】本発明は、個人が関心（興味や知見等も含む）をもつ知識領域を表す語句（以下、固有語と呼ぶ）を当該個人が書いたテキストデータから抽出して、当該個人の関心事を把握するための辞書を作成する装置に関する。

【０００２】

【従来の技術】最近、例えば、会議への出席者の選定やメールの宛先の選定等に供するために、個性や特徴といった個人の固有性データの必要性が増している。近年、日本においても、以前に比べて個人主義的な傾向が強まりつつある。これにより、各個人の個性が求められるようになりつつあり、また、他人と異なる個性が積極的に活用される場が増えつつある。

【０００３】一方、従来は様々なコミュニケーションが物理的な距離に大きく制約されていたが、ネットワーク等の発達により、必ずしも物理的に近くない人とのコミュニケーションや協業の可能性が大きくなってきた。これらにより、例えば、ある問題を解決するのに最もふさわしい人を広範囲に探したり、自分と興味が合う人を探して知り合いになるといった広範囲での人選の必要性が感じられるようになってきた。

【０００４】これに対して、従来では、例えば、何か解決したい問題が生じた場合には、ネットニュースやメイリング・リストを通じて不特定多数に呼びかけるという方法や、ネットニュース上に投稿されている記事やホームページの内容から判断して、問題を解決してくれそうだと思われる人に対して電子メイルなどを送り、解決を依頼するというような、ユーザ主導の方法がとられている。

【０００５】しかしながら、前者の場合には、不特定多数の相手に対して呼びかけるため、呼びかけられる相手側には自分に依頼されているという意識が少なく、呼びかけに答えるかどうかは相手の善意等に大きく依存する上に、本当にふさわしい人が誰かは呼びかける側からはわからない。また、後者の場合には、特定の個人に対して依頼されるため、前者に比べれば依頼に答えてもらえる可能性は高いが、現状では、偶然見たその人の投稿やホームページといった偶然性の強い情報に頼っているため、本当にふさわしい人を見つけるのは困難である。なお、この場合に、検索エンジン等を利用してホームページを検索して探すこともできるが、実際に内容を全部読んだりしない限り、ふさわしい人を見つけるのは困難であり、また、ふさわしい人を見つけるのにはユーザの負

担が非常に大きい。

【０００６】このため、近年、ある領域に関心や知見をもつ人を探したり、興味が一致する人同士を紹介したりするようなシステムが考えられ始めている。このようなシステムにおいて、人選がうまく行われるためには、各個人がどういう分野に知識や興味を持っているかといった、知識領域の固有性のデータが必要になってくる。なお、ここでの知識領域という語は、必ずしも学問の領域のような純粋に知的なものだけを指すのではなく、趣味などに関するものや、日常生活上の関心領域など、さまざまなものを指して用いている。

【０００７】スペシャリストや興味を同じくするグループの紹介といった、一種の人選を前提とした興味領域の表現手法の従来技術としては、特開平９－４４４７０号公報に開示されているものがある。この発明は、興味領域とその領域での情報処理能力を同時に表現しようとするものである。この中で興味領域の表現に着目すると、特定の人の興味領域はキーワードで表されており、そのキーワードは、その人が送受信した電子メイルのコンテンツから抽出されている。抽出方法は、やり取りされている電子メイル全体（その人が関係しないものも含む）における単語の出現頻度に対する特定の人が送受信した電子メイルにおける単語の出現頻度の比に基づくもので、従来の情報検索のためのキーワード抽出の手法を用いたものである。また、「分散型人脈活用支援システムにおける人脈データベースの構築」（重点領域研究「高度データベース」松江ワークショップ講演論文集Vol.1, pp109-114 (1996)）では、二人の間でやりとりされた電子メイルのサブジェクトに含まれる語を、二人に共通の興味領域として抽出する技術が開示されている。このように、従来においては、個人の関心領域は、単独のキーワード群で表されているだけであった。

【０００８】

【発明が解決しようとする課題】ところが、個人に対して、自分の関心領域を表す語（以下、固有語と呼ぶ）を自由に好きなだけ列挙してもらう実験を行ったところ、その結果から次のような特徴があることがわかった。

（１）例えば固有語が５０個列挙されている場合、その人は５０個の独立した領域に対して関心を持っているわけではなく、関心を持っている領域は幾つかの領域に分けられる。

（２）ひとつひとつの関心領域は、かなり抽象的なものから、非常に具体的なものまで、異なる抽象レベルの語で構成される。

（３）具体的な語としては、人名、地名、小説や映画の名前、商品名など、固有名詞が多く挙げられている。

（４）ひとつの領域を構成する語は互いに関連を持つが、その関係はシソーラスに表されるような厳密な上下関係というわけではなく、様々な意味、コンテキストに基づく関係をもつ。また、それはしばしば、その個人の経験や価値観に基づく、その人独自のものである。

【０００９】例えば、ある個人が挙げた固有語のうち、「文学」の領域に関連するものを、互いの関係を元に図式化したものを図８に示す。すなわち、単に「文学」という領域でも「小説」と「国文学」という異なる関心領域があり、その「小説」の中でも「推理小説」と「少女小説」との領域に分けられる。更に、「谷崎潤一郎」や「横溝正史」等といった具体的な関心対象は各領域に対して別個或いは重複した関連を有している。

【００１０】ディスカッションや問題解決において誰がふさわしいかを選ぶ場合、「どういう領域に関心があるか」という抽象的な領域と、その中で実際にどういうところに詳しかったり関心が強かったりするのかといった具体的な対象が共に必要である。抽象的な領域で興味が一致している場合でも、具体的な関心対象はまったく異なっていて、ぜんぜん話が合わないという場合もある。また、例えば、具体的な興味の対象として例えば同じ小説を挙げていたとしても、一方は文学に興味がありその中で特にその小説が好きであり、もう一方は文学には関心がないが、アフリカに興味があり、その小説がアフリカを描いているために関心を持っている、というような場合もある。このように、自分の関心を抽象度の違う語で表すことにより、お互いの意味を補強し合うことになるため、個人の関心領域をより正確に表すには、固有語は、単独のキーワードではなく、抽象度の異なるキーワードの組で表す必要がある。

【００１１】しかしながら、このような固有語を、ユーザ自身に列挙してもらうことは、ユーザにとって大きな負担となり、その場ですべてを思いつくことはかなり困難であり、また、抜けが多くなる。また、列挙してもらったとしても、列挙してもらった時の関心などに大きく左右される結果になり、かなり恣意的なものとなってしまう。さらに、関連する語句をまとめることも、ユーザにとって大きな負担となる。上述の例では単純化して「文学」に関するものだけを示したが、具体的な関心事は、複数の領域にまたがっていたりするため、非常に煩雑である。

【００１２】ここで、このような語句の関連を表したものとしてシソーラスがあり、シソーラスを利用した自動的な関連付けや階層化が考えられる。しかしながら、シソーラスは基本的に語の上位－下位関係を表したものであるが、固有語間の関係は上下関係ではないものが多いため適用するに適していない。さらに、現在、入手できる既存のシソーラスでは、固有名詞は一部の地名など以外記載されておらず、例え記載されているとしても、商品名や番組名のような日々増えていく語句に対してサポートされることは難しい。

【００１３】上記のような事情から、固有語の組をユーザに負担をかけることなく、自動的に抽出する方法が必要である。そこで、実際には、図８に示したように、数

段階の階層関係を持つものもあるが、まず、特に重要である最も下位にある具体的な固有語と、それより上の階層にあるより抽象的な固有語という２段階の組の抽出を考える。抽出する対象として、一般的にある人の関心が表われているものとして、そのユーザが作成したテキスト情報が考えられる。ここでは、洩れを減らすため、論文のようなオフィシャルなものから電子メイルのような日常的、プライヴェートなものまで含む種々のテキストで、かつある程度長い期間の間に書かれた、大量のテキストを想定する。

【００１４】従来、互いに関連を持つ語（以下、関連語と呼ぶ）をテキストから自動的に抽出する方法としては、語句の共出現頻度に基づくものがある。特開平６－１６８２７２号公報には、あるキーワードに対する関連語を作成するために、もとのキーワードを含む文書内に現れる他の語のうち、文書データベース全体における出現頻度に対して、もとのキーワードを含む文書内に現れる頻度の高いものをもとのキーワードの関連語とする発明が開示されている。このような方法によれば、シソーラスにはない上下関係以外の関連語も抽出することができる。しかしながら、このような関連語抽出方法は、情報検索において、指定したキーワードを含む大量の文書において、どのような関連語が使われているかを提示して、ユーザが結果を絞り込めるようにするためのものであり、結果の絞り込みにとっては有効な関連語でも、抽象度の異なる語の組といった点を考慮したものではなかった。

【００１５】本発明は、上記の事情に鑑みてなされたものであり、個人の知識領域の固有性を表す語句の組を正確に抽出し、これを辞書に作成する個人関心事辞書作成装置を提供することを目的としている。

【００１６】

【課題を解決するための手段】本発明に係る個人関心事辞書作成装置は、文書中から当該文書を作成した特定個人の具体的な関心を表す語を抽出して、辞書にまとめる個人関心事辞書作成装置であって、情報記憶手段に記憶されている特定個人によって作成されたテキストデータを含む文書を、テキスト解析手段で形態素解析等して語句の単位に切り分け、共起解析手段によって、テキストデータ中の予め定めたテキスト単位（例えば、文書毎や段落等）中で固有名詞と共起する語を当該テキストデータを作成した特定個人の知識領域を表す抽象固有語候補として抽出し、更に、当該固有名詞を抽出して各抽象固有語候補との共起頻度をカウントする。そして、この共起する固有名詞の共起頻度を基準として、抽象固有語抽出手段で抽象固有語候補から抽象固有語を抽出し、更に、具体固有語抽出手段により、抽出された抽象固有語と共起する固有名詞を、当該抽象固有語が表す領域での前記特定個人の具体的な関心を表す具体固有語として抽出し、これら抽出された抽象固有語と具体固有語とを対応付けて固有語辞書手段に保持する。

【００１７】また、本発明に係る個人関心事辞書作成装置では、共起解析手段が、テキストデータ中の予め定めたテキスト単位中で固有名詞と共起する語を当該テキストデータを作成した特定個人の知識領域を表す抽象固有語候補として抽出し、更に、当該固有名詞を抽出して個数をカウントし、この抽出された固有名詞数を基準として、抽象固有語抽出手段で抽象固有語候補から抽象固有語を抽出し、更に、具体固有語抽出手段により、抽出された抽象固有語と共起する固有名詞を、当該抽象固有語が表す領域での前記特定個人の具体的な関心を表す具体固有語として抽出し、これら抽出された抽象固有語と具体固有語とを対応付けて固有語辞書手段に保持する。

【００１８】このようにして作成された固有語辞書中の固有語の組は、その個人にとっての関連語として捉えることができ、したがって、人選等における利用のみならず、例えば、情報検索等においても、関連語情報としてその個人に合わせた検索の支援に用いることができる。例えば、情報検索における関連語の利用には、クエリーの拡張に用いられる場合と、結果の絞り込みに用いられる場合とがある。クエリー拡張の場合は、関連語によってクエリーを拡張して検索することにより、より多くの結果を得る。一方、結果の絞り込みの場合は、クエリーの関連語としてどのような語が使われているかによって、ユーザの検索意図に合ったものだけに絞り込む。

【００１９】本発明で抽出される固有語は、この両方の用途に用いることができる。前者の場合には、固有名詞のような具体的な語は、意味の特定性が高いため、クエリーの関連語である固有名詞が含まれている文書を検索することにより、クエリーの語が含まれていなくても必要な文書を検索することができる。また、後者の場合には、クエリーを満たす文書のうち、どのような固有名詞が共起しているかによって、ユーザの意図する文書だけを検索することができる。更に、本発明で抽出される固有語は、この他にも、個人の関心領域の情報を必要とする、個人の知的な活動を支援するようなあらゆるシステムにおいて利用しうる。

【００２０】

【発明の実施の形態】本発明の一実施形態を図面を参照して説明する。図１には、本実施形態に係る個人関心事辞書作成装置の構成を示してあり、この個人関心事辞書作成装置は、情報記憶部１、テキスト解析部２、共起解析部３、抽象固有語抽出部４、具体固有語抽出部５、固有語登録部６、および固有語辞書７を備えている。なお、本実施形態における各機能手段２～６および後述する各機能手段２１、２２、３１、３２は、予めインストールしたプログラムをコンピュータハードウエア資源を用いて実行することにより構成される。

【００２１】情報記憶部１は、例えば、磁気ディスク装置等といった情報を読み書き自在に記憶する装置を有

し、特定個人によって書かれたテキストデータを含む文書を複数記憶する。なお、情報記憶部1は、テキスト解析部2等からネットワーク上の離れた場所に設置してもよく、また、複数の装置によって構成してもよい。また、上記した文書にはテキストデータの他に絵や図形などのデータを含んでいてもよい。

【0022】テキスト解析部2は、情報記憶部1に記憶されている文書中のテキストデータを解析し、テキストデータ中の自立語（名詞、形容詞、動詞等）を切り出す処理を行う。なお、本実施形態では、テキスト解析部2は形態素解析部21と語処理部22とを備えており、テキストデータ中から名詞を切り出す処理を行う。形態素解析部21は、情報記憶部1に記憶されている文書中のテキストデータに対して形態素解析を行うことによって単語に分割し、各単語に品詞情報を付与する処理を行う。なお、形態素解析は、自然言語処理の基礎技術として広く知られており、例えば「自然言語処理の基礎技術」（野村浩郷著、社団法人電子情報通信学会、1988）や「情報処理」（Vol.30, No.10, 1989）の「3．1形態論」等に記載されている方法により、容易に実現することができる。

【0023】語処理部22は、形態素解析部21の解析結果に基づき、未知語と複合語の処理を行う。例えば、固有名詞には外来語や新語が多いため、未知語を固有名詞とみなし、固有名詞の品詞情報を与える。また、漢字の名詞の連続、カタカナの名詞の連続は複合語と考えられるので、まとめてひとつの名詞として品詞情報を与える。なお、本実施形態では、既存の形態素解析システムを用いることを前提として上述のような構成としたが、本発明はこれに限らず、形態素解析部21と語処理部22の処理を同時に行うようにしてもよい。この際、予め定めたルールに基づいて、例えば、「　」や“　”などで囲まれた短い語句を固有名詞とするようにしてもよい。

【0024】また、テキスト解析部2に、形態素解析用の辞書とは別に固有名詞辞書を設け、未知語や複合語をこの辞書中の固有名詞とマッチングして、マッチするものは固有名詞とするようにしてもよい。また、本実施形態では、テキスト解析部2の解析内容として形態素解析を用いたが、本発明はこれに限らず、共起解析部3で必要とされる情報にしたがって、構文解析など、さらに深い解析を行うようにしてもよい。

【0025】共起解析部3は、あるテキスト単位毎に、固有名詞と共起する語を文書データの作成者の知識領域を表す抽象固有語候補として抽出し、それぞれの抽象固有語候補に対して、共起する固有名詞の共起頻度をカウントする。ここで、本実施形態において、「共起する」とは、「同一テキスト単位中に共存する」ことを指すものとするが、本発明はこれに限らず、例えば、文法的な係り受け関係をもつ場合を指すようにしてもよい。なお、この場合には、テキスト解析部2において、構文解析等のさらに深い解析を行うようにすればよい。

【0026】また、テキスト単位とは、本実施形態では一文書としており、同じ文書内に共に存在していれば「共起する」とみなす。なお、本発明はこれに限らず、予め定めたテキスト単位であればどのような単位でもよく、例えば、一文、一段落などをテキスト単位としてもよい。また、情報記憶部1に記憶された文書が構造化文書である場合には、特定のタグによって囲まれた範囲としてもよい。また、構造的な特徴以外に、一定の行数、文字数などとしてもよい。文法的な係り受け関係をもつ場合に「共起する」とする場合には、テキスト単位としては一文を用いるのが好ましい。

【0027】また、固有名詞と共起する語として、本実施形態では名詞を考えており、ここでは、サ変動詞語幹もサ変名詞とみなして、名詞の中に含め、一方、形式名詞と固有名詞は除く。以後、本実施形態中の抽象固有語候補について、断りなく名詞と書いた場合には、上述の範囲の名詞を指すものとする。なお、本発明はこれに限らず、要は、個人の固有語となり得る語であれば何でもよく、例えば、広く自立語全般、名詞と動詞全部などとしてもよく、また、サ変動詞語幹などを含まない名詞のみとしてもよく、また、目的に応じて形容詞や形容動詞などとしてもよい。

【0028】本実施形態では、共起解析部3は、単語抽出部31と共起カウント部32とを備えている。単語抽出部31は、一文書毎に、文書中の固有名詞と名詞を、それぞれ重複を除いて抽出し、文書毎の固有名詞リストおよび抽象固有語候補リストを作成する。ここで、重複を除くとは、例えば固有名詞リストの作成において、同じ固有名詞が同一文書中に2回以上出てきた場合、2回目以降は無視するということである。共起カウント部32は、全文書の固有名詞リストおよび抽象固有語候補リストから、各抽象固有語候補毎に、共起する固有名詞の共起頻度（本実施形態の場合、共出現文書数）をカウントし、共起リストを作成する。

【0029】抽象固有語抽出部4は、共起する各固有名詞の共起頻度に基づいて、共起リストの抽象固有語候補の中から、文書の作成者固有の知識領域を表す語を選択する。この選択の基準は、単純には、共起する固有名詞が何個以上で、トータルの出現文書数がいくつ以上であるといったものである。また、これらの数値を含む評価式を作り、その評価値が予め決めた値以上になるもの、あるいは、評価値が大きい方から何%までといった基準でもよい。また、後述するように共起する固有名詞の個数や、各固有名詞の出現文書数以外の要素をさらに加味してもよく、要は、共起する固有名詞の個数および各固有名詞の出現文書数に基づくものならどのような方法でもよい。

【0030】具体固有語抽出部5は、抽象固有語抽出部

４で抽出された各抽象固有語に関連する具体固有語を選択する。ここでは、選択の基準は、例えば、共起リストにある固有名詞のうち、共起頻度が予め決めた数より多いものを選択する。この選択の基準は、この他に、共起文書数が多いものからｎ個選ぶ、全体のｎ％選ぶなどでもよく、また、共起頻度以外の基準を組み合わせて選択するようにしてもよい。

【００３１】固有語登録部６は、抽出された抽象固有語と具体固有語をセットにして、文書作成者の固有語辞書７に登録する処理を行う。ここで、各固有名詞の頻度や、抽象固有語抽出部４、具体固有語抽出部５で算出した評価値などを共に登録するようにしてもよい。固有語辞書７は、例えば、磁気ディスク装置等といった情報を読み書き自在に記憶する装置で構成されており、上記の抽象固有語と具体固有語とのセットを個人毎の辞書として記憶する。

【００３２】次に、上記の個人関心事辞書作成装置によって行う、特定個人に対する固有語辞書を作成する処理を図２～図７を参照して説明する。なお、情報記憶部１には、Ａ氏によって作成されたＮ個の文書が記憶されているものとする。まず、テキスト解析部２の形態素解析部２１において、情報記憶部１に記憶されている全文書中のテキストデータを形態素解析する（ステップＳ１）。すなわち、形態素解析部２１は、まず、ある１つの文書Ｄ１からテキストデータを一文ずつ読み込んで単語に分割し、各単語に品詞情報を付与する。例えば、テキスト中に「ベナンのチャイナタウンは、伝統的な景観がかなり残されている。」という一文があった場合には、形態素解析によって図３に示すように、当該一文を単語に切り分けてその品詞情報を付与する。このように文書Ｄ１のすべての文の解析が終了すると、形態素解析部２１は次の文書Ｄ２に移って同様の解析を行う。

【００３３】そして、全文書Ｎについて形態素解析が終了すると、語処理部２２において、未知語と複合語の処理を行う（ステップＳ２）。例えば、図３に示した解析結果が得られた場合、語処理部２２の処理が行われると、図４に示すように、未知語である「ベナン」が固有名詞に、名詞の連続である「チャイナ」「タウン」がひとつにまとめられて「チャイナタウン」という名詞になる。なお、未知語、複合語の処理は、文書Ｄ１の先頭から順番に文書ＤＮまで行われる。

【００３４】次いで、共起解析部３の単語抽出部３１において、固有名詞リストおよび抽象固有語候補リストが作成される（ステップＳ３）。単語抽出部３１は、まず最初の文書Ｄ１からすべての固有名詞と名詞を重複なく抽出し、固有名詞リストＰＮＬ１および抽象固有語候補リストＡＩＬ１を作成する。例えば、「ベナンのチャイナタウンは、伝統的な景観がかなり残されている。ジョージタウンの商店建築は、ショップハウスと呼ばれる、住居と商店を兼ねた形式である。」というテキスト文章のみからなる文書があった場合、固有名詞リストには「ベナン」「ジョージタウン」という固有名詞が載せられ、抽象固有語候補リストには「「チャイナタウン」「伝統」「景観」「商店建築」「ショップハウス」「住居」「商店」「形式」という語が載せられる。このように文書Ｄ１についてリストを作成すると、単語抽出部３１は次の文書Ｄ２に移って同様にリストを作成し、すべての文書Ｎについて、各リストを作成する。

【００３５】次いで、共起カウント部３２において、抽象固有語候補毎に、共起する固有名詞をカウントし、共起リストＣＯＬを作成する（ステップＳ４）。この共起カウント部３２による動作を図５を参照して詳しく説明すると、共起カウント部３２は、まず、最初の文書Ｄ１を取り出し（ステップＳ１１、Ｓ１２）、当該文書Ｄ１の中小固有語候補リストに単語があるか否かを確認し（ステップＳ１３）、単語がない場合には次の文書Ｄ２への処理へ移行する一方、単語がある場合には文書Ｄ１の抽象固有語候補リストＡＩＬ１の先頭から一語（Ｗ）取り出す（ステップＳ１４）。

【００３６】次いで、取り出した語Ｗが共起リストＣＯＬにあるかを調べ（ステップＳ１５）、ない場合には当該語をリストに追加する一方（ステップＳ１６）、ある場合には文書Ｄ１の固有名詞リストＰＮＬ１に単語があるかを確認し（ステップＳ１７）、単語がない場合には次の抽象固有語への処理へ移行する一方、単語がある場合には文書Ｄ１の固有名詞リストＰＮＬ１の先頭から一語（ＰＮ）取り出す（ステップＳ１８）。そして、この語（ＰＮ）が共起リスト中のＷの項の共起固有名詞中にあるかどうかを調べ（ステップＳ１９）、ない場合にはＰＮをリストに追加してその頻度を１とする一方（ステップＳ２０）、ある場合には、その頻度を＋１加算する（ステップＳ２１）。

【００３７】上記のような処理によって、固有名詞リストの全固有名詞についてリストの変更を終わると、抽象固有語候補リストの次の語に移り、同様に固有名詞リストの先頭の語から順に、共起リストへの追加または頻度の加算を行う。更に、文書Ｄ１の抽象固有語候補リストのすべての名詞についてこの処理が終了すると、次の文書Ｄ２に移り、文書Ｄ２の抽象固有語候補リストＡＩＬ２と固有名詞リストＰＮＬ２について同様の処理を行う。以後同様に、文書Ｎまでの抽象固有語候補リストおよび固有名詞リストを処理し、共起リストＣＯＬを更新する。例えば、文書Ｄ１と文書Ｄ２の抽象固有語候補リストと固有名詞リストが図６（ａ）に示すものであった場合、文書Ｄ１と文書Ｄ２についての処理を終わった時点での共起リストＣＯＬの一部を示すと図６（ｂ）のようになる。なお、図６（ｂ）中に示す数字は頻度である。

【００３８】上記のようにして全文書Ｎに対する共起リストＣＯＬが作成されると（ステップＳ４）、抽象固有

11

語抽出部４において、各固有名詞の出現文書数に基づいて、共起リスト中の抽象固有語候補の中から文書の作成者固有の知識領域を表す語を選択する（ステップＳ５）。そして、抽象固有語の抽出が終了すると、具体固有語抽出部５において、選択された各抽象固有語に関連する具体固有語を抽出する（ステップＳ６）。ここでは、共起リストにある固有名詞のうち、共起頻度が予め決めた数より多いものを選択する。例えば、図７（ａ）に一部を示すような共起リストＣＯＬが作成された場合、例えば、しきい値を４０として、図７（ｂ）に示すような抽象固有語と具体固有語の組が、抽象固有語抽出部４、具体固有語抽出部５の動作により抽出される。

【００３９】このようにして具体固有語の選択が終了すると、固有語登録部６において、抽出された抽象固有語と具体固有語をセットにして、文書の作成者の固有語辞書７に登録し（ステップＳ７）、一連の辞書作成動作が終了する。このようにして作成された固有語辞書７は、例えば、個人の関心や知見に基づいて問題を解決してくれる人を選んだり、興味の一致する人を紹介したりする一種の人選システムにおいて、個人の関心領域を表す有効な情報として利用することができる。なお、本実施形態では本発明の個人関心事辞書作成装置を、情報を永続的に記憶保持する辞書を作成するものとして説明したが、この辞書は情報を一時的に保持するバッファ的なものとしてもよく、例えば、固有語を抽出した結果を永続的な情報としては登録せずに、アプリケーション中で直接的に利用するようにしてもよい。

【００４０】上記した実施形態では、共起カウント部３２は、固有名詞リストおよび抽象固有語候補リストから共起する固有名詞の共起頻度をカウントして共起リストを作成し、抽象固有語抽出部４は、各固有名詞の共起頻度に基づいて、共起リストの抽象固有語候補の中から文書の作成者固有の知識領域を表す語を選択するようにしたが、本発明の他の実施態様として、共起カウント部３２は、全文書の固有名詞リストおよび抽象固有語候補リストから、各抽象固有語候補毎に、共起する固有名詞の個数をカウントして共起リストを作成し、抽象固有語抽出部４は、共起する固有名詞の個数に基づいて、共起リスト

12

＊の抽象固有語候補の中から、文書の作成者固有の知識領域を表す語を選択するようにしてもよい。すなわち、抽象固有語に関連する固有名詞の数によって、登録のために選択する抽象固有語を選択するようにしてもよい。

【００４１】

【発明の効果】以上説明したように、本発明によると、固有名詞と共起する抽象固有語候補の中から、その固有名詞の個数や共起頻度に基づいて、特定個人の具体的な関心事を表す具体固有語として選定するようにしたため、個人の知識領域の固有性を表す語句の組を正確に抽出し、これを辞書に作成することができる。このため、この辞書を用いて、例えば、人選やメールの宛先選定を適切に行うことができ、個人の知的活動に係わる業務を円滑に実行することができる。

【図面の簡単な説明】

【図１】　本発明の一実施形態に係る個人関心事辞書作成装置の構成図である。

【図２】　本発明の一実施形態に係る処理動作を説明するフローチャートである。

【図３】　本発明の一実施形態に係るテキスト解析部の動作を説明する図である。

【図４】　本発明の一実施形態に係るテキスト解析部の動作を説明する図である。

【図５】　本発明の一実施形態に係る共起カウント部の動作を説明するフローチャートである。

【図６】　本発明の一実施形態に係る共起カウント部の動作を説明する図である。

【図７】　本発明の一実施形態に係る抽象固有語抽出部および具体固有語抽出部の動作を説明する図である。

【図８】　従来の課題を説明する図である。

１・・・情報記憶部、　２・・・テキスト解析部、　３・・・共起解析部、４・・・抽象固有語抽出部、　５・・・具体固有語抽出部、５・・・固有語登録部、　７・・・固有語辞書、　２１・・・形態素解析部、２２・・・語処理部、　３１・・・単語抽出部、　３２・・・共起カウント部、

【図３】

ベナン	の	チャイナ	タウン	は	伝統
未知語	助詞	名詞	名詞	助詞	名詞

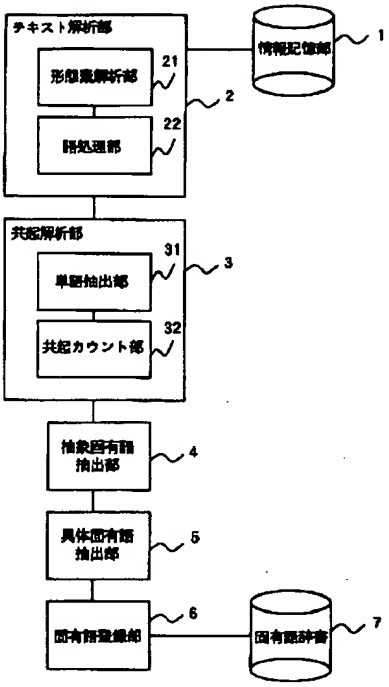
的だ	最低	が	かなり	残す	れる	て	いる。
形容動詞性接尾辞	名詞	助詞	副詞	動詞	助動詞	助動詞	助動詞

【図４】

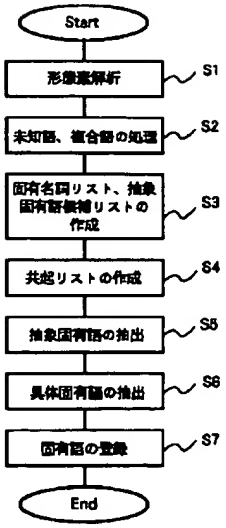
ベナン	の	チャイナタウン	は	伝統
固有名詞	助詞	名詞	助詞	名詞

的だ	最低	が	かなり	残す	れる	て	いる。
形容動詞性接尾辞	名詞	助詞	副詞	動詞	助動詞	助動詞	助動詞

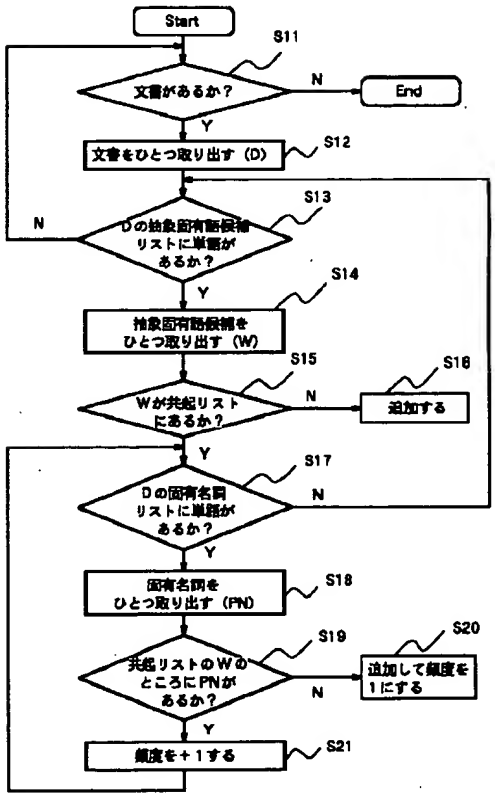
【図１】


テキスト解析部
形態素解析部
21
2
語処理部
22
情報記憶部
1
共起解析部
単語抽出部
31
3
共起カウント部
32
抽象固有語抽出部
4
具体固有語抽出部
5
固有語登録部
6
固有語辞書
7

【図２】


Start
形態素解析
S1
未知語、複合語の処理
S2
固有名詞リスト、抽象固有語候補リストの作成
S3
共起リストの作成
S4
抽象固有語の抽出
S5
具体固有語の抽出
S6
固有語の登録
S7
End

【図５】


Start
S11
文書があるか？
N
End
Y
文書をひとつ取り出す（D）
S12
S13
Dの抽象固有語候補リストに単語があるか？
N
Y
S14
抽象固有語候補をひとつ取り出す（W）
S15
Wが共起リストにあるか？
N
S16
追加する
Y
S17
Dの固有名詞リストに単語があるか？
N
Y
固有名詞をひとつ取り出す（PN）
S18
S19
共起リストのWのところにPNがあるか？
N
S20
追加して頻度を１にする
Y
頻度を＋１する
S21

【図6】

(a)

文書D1		文書D2	
抽象固有語 候補リストAIL1	固有名詞 リストPNL1	抽象固有語 候補リストAIL2	固有名詞 リストPNL2
チャイナタウン	ペナン	建築	パリ
景観	ジョージタウン	都市計画	ジョージタウン
建築		チャイナタウン	
伝統			
ショップハウス			

(b) 共起リストCOL

チャイナタウン		景観		建築		伝統		…
ペナン	1	ペナン	1	ペナン	1	ペナン	1	…
ジョージタウン	2	ジョージタウン	1	ジョージタウン	2	ジョージタウン	1	…
パリ	1			パリ	1			…

【図7】

(a)

野球		音楽		映画		特許		…
阪神タイガース	85	春の振鳥	92	小津安二郎	72	A社	5	…
イチロー	77	ピチカート・ファイヴ	88	台湾	66	B社	3	…
掛布	42	××ゲーム	6	○○牛乳	7			…
…		…		…				…

(b)

抽象固有語	野球	音楽	映画
具体固有語	阪神タイガース イチロー 掛布	春の振鳥 ピチカート・ファイヴ	小津安二郎 台湾

【図8】